Journal of Computer-Aided Molecular Design, **16**: 11–26, 2002.
KLUWER/ESCOM

Further development and validation of empirical scoring functions for structure-based binding affinity prediction

Renxiao Wang[a], Luhua Lai[b] & Shaomeng Wang[a,*]
[a]*Medical Chemistry and Comprehensive Cancer Center, University of Michigan, 1500 E. Medical Center Drive, Ann Arbor, MI 48109-0934, U.S.A.;* [b]*Institute of Physical Chemistry, Peking University, Beijing 100871, P.R. China*

Received 27 August 2001; Accepted 7 February 2002

Key words: binding affinity prediction, consensus scoring, empirical scoring molecular docking, structure-based drug design

Summary

New empirical scoring functions have been developed to estimate the binding affinity of a given protein-ligand complex with known three-dimensional structure. These scoring functions include terms accounting for van der Waals interaction, hydrogen bonding, deformation penalty, and hydrophobic effect. A special feature is that three different algorithms have been implemented to calculate the hydrophobic effect term, which results in three parallel scoring functions. All three scoring functions are calibrated through multivariate regression analysis of a set of 200 protein-ligand complexes and they reproduce the binding free energies of the entire training set with standard deviations of 2.2 kcal/mol, 2.1 kcal/mol, and 2.0 kcal/mol, respectively. These three scoring functions are further combined into a consensus scoring function, X-CSCORE. When tested on an independent set of 30 protein-ligand complexes, X-CSCORE is able to predict their binding free energies with a standard deviation of 2.2 kcal/mol. The potential application of X-CSCORE to molecular docking is also investigated. Our results show that this consensus scoring function improves the docking accuracy considerably when compared to the conventional force field computation used for molecular docking.

Introduction

Considerable advances in structure-based drug design have made a significant impact on drug discovery processes in the past decade [1–5]. By utilizing the essential structural properties of the target macromolecule, a variety of methods now exist for suggesting potential ligand molecules either by screening large chemical databases [6–10] or by assembling molecular fragments inside the binding site [11–18]. These methods usually suggest a large number of molecules rapidly, far too many for organic synthesis and biological experiments. Therefore, a structure-based drug design approach tends to arrive at the bottleneck where it is necessary to select only the most promising candidates for further experimental characterization. The basic assumption underlying structure-based drug design is that a good ligand molecule should bind tightly to its target. Thus, it is extremely valuable to predict the binding affinity of a given ligand to its target and use it as a criterion for selection. This is known as the 'scoring problem' and has attracted great interests in developing methods for binding affinity calculation [19–21].

A large group of methods calculate binding affinities through force fields. In early years, attempts have been made to calculate the direct interactions, e.g. steric and electrostatic interactions, between a ligand and its target molecule and relate the force field energies to binding affinities [22]. This method is still popular nowadays especially among molecular docking studies. However, as many researchers have pointed out, the interaction energy computed in this

*To whom correspondence should be addressed. E-mail: shaomeng@med.umich.edu

way is only an approximation to the enthalpy change in the binding process, therefore the application of this method is usually restricted to the analysis of a congeneric series of ligands. Some researchers have supplemented standard force fields with an additional term to address the solvation effect with either PB/SA or GB/SA method [23]. More ambitious methods, such as free energy perturbation [24] and linear response approximation [25, 26], try to consider solvent molecules explicitly and deal with ensemble averages. In theory these methods are expected to give more accurate predictions. However, in practice they do not always meet this expectation due to the deficiency in the force field as well as in the sampling procedure. In addition, these methods are still computationally expensive even for today's computers, which has limited their popularity in structure-based drug design practice.

Following the pioneering work of Böhm [27], a number of so-called empirical scoring functions have emerged as an alternative [28–32]. These approaches assume that the overall receptor-ligand binding free energy can be decomposed into basic components, which can be written out conceptually as:

$$\Delta G_{\text{bind}} = \Delta G_{\text{motion}} + \Delta G_{\text{interaction}} + \Delta G_{\text{desolvation}} + \Delta G_{\text{configuration}}.$$

Usually those factors which are known to be important for the binding process are included in the above function. Unlike force fields, empirical scoring functions are not derived from 'first principle'. Instead, they are directly calibrated with a set of protein-ligand complexes with experimentally determined structures and binding affinities through multivariate regression analysis. Empirical scoring functions have several appealing features. Firstly, since they are calibrated with diverse protein-ligand complexes, their applications are not limited to a certain congeneric series of ligands or a particular target receptor. Secondly, each term in an empirical scoring function has a clear physical meaning. Studying the regression coefficients before each term sheds lights on the understanding of the receptor-ligand binding process. Thirdly, at a lightning speed, the accuracy level (~2 kcal/mol) that a current empirical scoring function can achieve in binding affinity prediction is acceptable for structure-based drug design approaches. In recent years, empirical scoring functions have become more and more popular among structure-based drug design applications in which very accurate binding affinity predictions are not necessary, such as virtual database screening and *de novo* ligand generation.

We have extensive experience in applying several empirical scoring functions, including Bohm's scoring function [27], ChemScore [30] and SCORE [32], to structure-based drug design projects. Despite of all the encouraging results we have obtained with these empirical scoring functions, it is clear that there is still plenty of room for improvement in terms of accuracy as well as robustness. In this paper, we will describe our work on further development and validation of empirical scoring functions. Firstly, we have derived three scoring functions, each of which has only five adjustable parameters. These scoring functions are calibrated with a diverse set of 200 protein-ligand complexes, which is the largest one ever used by an empirical scoring function approach. Secondly, inspired by the consensus scoring strategy [33], we combine these three scoring functions into a consensus scoring function, X-CSCORE, to ensure converged results in binding affinity prediction. This consensus scoring function is tested on an independent set of 30 protein-ligand complexes. Thirdly, we have also explored the potential application of X-CSCORE to molecular docking. When compared to conventional force field computation, this consensus scoring function performs considerably better in identifying the experimentally determined protein-ligand complex structures.

Methods and results

Training set construction

Developing an empirical scoring function requires a set of receptor-ligand complexes for calibration. Both the size and the quality of the training set will affect the final form of the scoring function. In our selection of receptor-ligand complexes, we used the following five criteria to ensure the quality of the training set. (1) Only protein-ligand complexes are considered. Complexes involving other types of receptors, such as nucleic acids, are not included. (2) The ligand molecule should be a 'normal' organic compound and bind to the receptor non-covalently. Therefore, complexes containing covalently bound ligands, complex ligands (such as Heme), or large ligands (MW > 1000) are excluded. (3) There should be no cofactor binding beside the ligand. (4) Crystal structure of the complex with a resolution better than 3.0 Å should be available from the Protein Data Bank (PDB) [34]. Complex

structures solved by NMR techniques are currently not included in our selection. (5) The dissociation equilibrium constant (K_i or K_d) of the complex has been determined experimentally and can be found in literature. Complexes with only IC_{50} values are not accepted.

The resulting training set has 200 protein-ligand complexes, which comprises more than 70 different types of proteins. Basically, this training set is an assembly of the training sets used by other empirical scoring functions [27–32] plus our own collections. The experimentally determined binding affinities are cited either from those previous approaches or the references listed in the relevant PDB files. All binding affinities are expressed in the negative logarithms of dissociation constants, i.e. pK_d, for convenience. In this training set, the pK_d values range from 1.48 to 11.42, covering nearly 10 orders of magnitude. Here we neglect the potential inconsistence in the dissociation constants related to experiment conditions, such as PH level, temperature, and salt concentration. A complete list of the training set can be found in the *supplementary material* section in this paper.

Coordinates of the complex structure in the training set are downloaded from PDB. No minimization is performed to further adjust the structure. For the convenience of processing, each complex structure is processed in SYBYL [35]. First, the ligand is extracted from the complex, assigned proper atom and bond types, and then written out as a separate file in the MOL2 format. The remaining part of the complex, i.e. the protein, is written out into another file in the PDB format. Metal ions located inside the binding site are left with the protein and treated as part of it. All crystallographic water molecules and other cofactors are removed.

Scoring functions

We assume that the overall free energy change in a protein-ligand binding process can be dissected into the following terms:

$$\Delta G_{\text{bind}} = \Delta G_{\text{vdw}} + \Delta G_{\text{H - bond}} + \Delta G_{\text{deformation}} + \Delta G_{\text{hydrophobic}} + \Delta G_0. \quad (1)$$

Here, ΔG_{vdw} accounts for the van der Waals interaction between the ligand and the protein; $\Delta G_{\text{H - bond}}$ accounts for the hydrogen bonding between the ligand and the protein; $\Delta G_{\text{deformation}}$ accounts for the deformation effect; $\Delta G_{\text{hydrophobic}}$ accounts for the hydrophobic effect; ΔG_0 is the regression constant which implicitly includes the effects due to the translational and rotational entropy loss in the binding process. Detailed algorithms for calculating each term will be described below.

(1) Atom classification. Besides element type and hybridization state, both ligand and protein atoms need to be classified to compute some of the terms in our scoring functions. The atom types defined in our study are: (i) H-bond donor. Oxygen and nitrogen atoms bonded to hydrogen atom(s) and metal ions located inside the binding site of the protein. (ii) H-bond acceptor. Oxygen and sp^2 or sp hybridized nitrogen atoms with lone pair(s). (iii) H-bond donor/acceptor. Oxygen and nitrogen atoms which may act as either H-bond donor or H-bond acceptor, such as the oxygen atom in a hydroxyl group. (iv) Polar atom. Oxygen and nitrogen atoms that are neither H-bond donor nor H-bond acceptor, sulfur and phosphorus atoms, and carbon atoms bonded to hetero-atom(s). (v) Hydrophobic atom. Carbon atoms that do not belong to the 'polar atom' group and halogen atoms.

The following set of atomic radii are used in computation: carbon, 1.9 Å; nitrogen, 1.8 Å; oxygen, 1.7 Å; sulfur, 2.0 Å; phosphorus, 2.1 Å; fluorine, 1.5 Å; chlorine, 1.8 Å; bromine, 2.0 Å; iodine, 2.2 Å; metals, 1.2 Å. This radii set is applied to both ligands and proteins.

(2) Van der Waals interaction. The van der Waals interaction is one of the essential non-covalent interactions. We employ the Lennard-Jones equation to reflect the balance between the short-range repulsion and the long-range attractive dispersion force:

$$VDW = \sum_{i}^{ligand} \sum_{j}^{protein} VDW_{ij} = \sum_{i}^{ligand} \sum_{j}^{protein} \left[\left(\frac{d_{ij,0}}{d_{ij}} \right)^8 - 2 \times \left(\frac{d_{ij,0}}{d_{ij}} \right)^4 \right] \quad (2)$$

Here *VDW* denotes for the van der Waals interaction energy, which is calculated by considering all the atom pairs between the ligand and the protein; d_{ij} denotes for the distance between the ligand atom i and the protein atom j; $d_{ij,0} = r_i + r_j$, i.e. the sum of van der Waals radius of atom i and j. Note that we use a 'softer' form in Equation 2 instead of the standard 12–6 equation. Furthermore, in our algorithm, (i) only heavy atoms contribute. Hydrogen atoms are neglected. (ii) All heavy atoms are weighted equally. No weight factor is used to differentiate them. (iii) To avoid the huge repulsion raised by overlapped atom

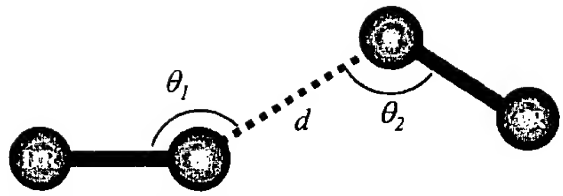


Figure 1. Illustration of the three geometric parameters used in characterizing a hydrogen bond.

pairs, we set an upper limit of 100 for VDW_{ij} in Equation 2. For any pair of atoms, if VDW_{ij} exceeds this limit, it will be cut flat to 100.

(3) Hydrogen bonding. Hydrogen bonding is perhaps the most important factor for the specific binding of a ligand to its receptor. Such interaction happens when two atoms get close enough and form a donor-acceptor pair. The geometry of a hydrogen bond, D–H· · ·A, is typically described by the bond length, i.e. the distance between the hydrogen atom (H) and the acceptor (A), and the bond angle, i.e. ∠DHA. However, hydrogen atoms are normally not revealed in X-ray crystallography analysis. Although hydrogen atoms can be added later, energy minimization is usually required to set them into position. This practice could become problematic especially when the hydrogen atoms can take multiple low-energy positions, such as the one in a hydroxyl group. Furthermore, minimized structures will depend on force field parameters and they may be incompatible with the initial experimentally determined ones. Therefore, we choose not to consider hydrogen atoms explicitly. Here we introduce the concept of 'root': the root of an atom is its non-hydrogen neighboring atom. When an atom bonds with more than one non-hydrogen atom, its root locates at the geometric center of all its non-hydrogen neighboring atoms. Let DR denotes for the donor's root and AR for the acceptor's root. In our algorithm, the geometry of a hydrogen bond is described by: (i) the distance (d) between D and A; (ii) the angle (θ_1) between DR, D and A; and (iii) the angle (θ_2) between D, A and AR (Figure 1).

We assume that a hydrogen bond has an ideal geometry and any deviation from it will weaken the strength of the hydrogen bond. The strength of a hydrogen bond is then computed by considering these three geometric descriptors:

$$HB_{ij} = f(d_{ij})\, f(\theta_{1.ij})\, f(\theta_{2.ij}). \qquad (3)$$

The distance function $f(d)$ and the angular functions $f(\theta_1)$ and $f(\theta_2)$ in Equation 3 are written in the following simple linear fuzzy forms:

$$\begin{aligned}
f(d) &= 1.0 && d_0 \le d_0 - 0.7\ \text{Å} \\
&= (1/0.7) \times (d_0 - d) && d_0 - 0.7\text{Å} < d \le d_0 \\
&= 0.0 && d > d_0 \\
f(\theta_1) &= 1.0 && \theta_1 \ge 120° \\
&= (1/60) \times (\theta_1 - 60) && 120° > \theta_1 \ge 60° \\
&= 0.0 && \theta_1 < 60° \\
f(\theta_2) &= 1.0 && \theta_2 \ge 120° \\
&= (1/60) \times (\theta_2 - 60) && 120° > \theta_2 \ge 60° \\
&= 0.0 && \theta_2 < 60°
\end{aligned}$$

Here $d_0 = r_i + r_j$, i.e. the van der Waals distance between the donor and the acceptor. These functions are derived from the analysis of all the potential hydrogen bonding pairs presented in the training set. The observed distribution of the donor-acceptor distance (d) is shown in Figure 2a. In this figure, one can see that the peak value appears around 2.8Å, which can be interpreted as the ideal length of a hydrogen bond. As d increases, the population decreases. But after d exceeds 3.4 ~ 3.5 Å, it passes the bottom and begins to rise again, which can be interpreted as the turning point from a hydrogen bond to a normal van der Waals contact. Therefore, it is reasonable to define that $f(d) = 1.0$ when $d = 2.8$ Å while $f(d) = 0.0$ when $d = 3.5$ Å. Considering the atomic radii of oxygen and nitrogen atoms, 2.8 Å corresponds to $d_0 - 0.7$ Å while 3.5 Å corresponds to d_0, approximately. By assuming that the distance dependence of the strength of a hydrogen bond is linear within this range, one will obtain the function listed above. The angular functions $f(\theta_1)$ and $f(\theta_2)$ are also derived similarly by interpreting the observed distributions of θ_1 and θ_2 from the training set (Figures 2b and 2c).

The hydrogen bonding interaction between the ligand and the protein is calculated by summing up all the hydrogen bonds:

$$HB = \sum_{i}^{ligand} \sum_{j}^{protein} HB_{ij} \qquad (4)$$

All types of hydrogen bonds, i.e. O-O, O-N, and N-N, are equally weighted so that no extra parameter is necessary. Special attention has been paid to the saturation in hydrogen bonding if one donor or acceptor atom contacts with multiple donor or acceptor atoms. For a given donor or acceptor atom, we define that (i) the maximal number of hydrogen bonds that a donor atom can form should not exceed the number of hydrogen atoms on that donor atom; and (ii) the maximal number of hydrogen bonds that an acceptor atom can

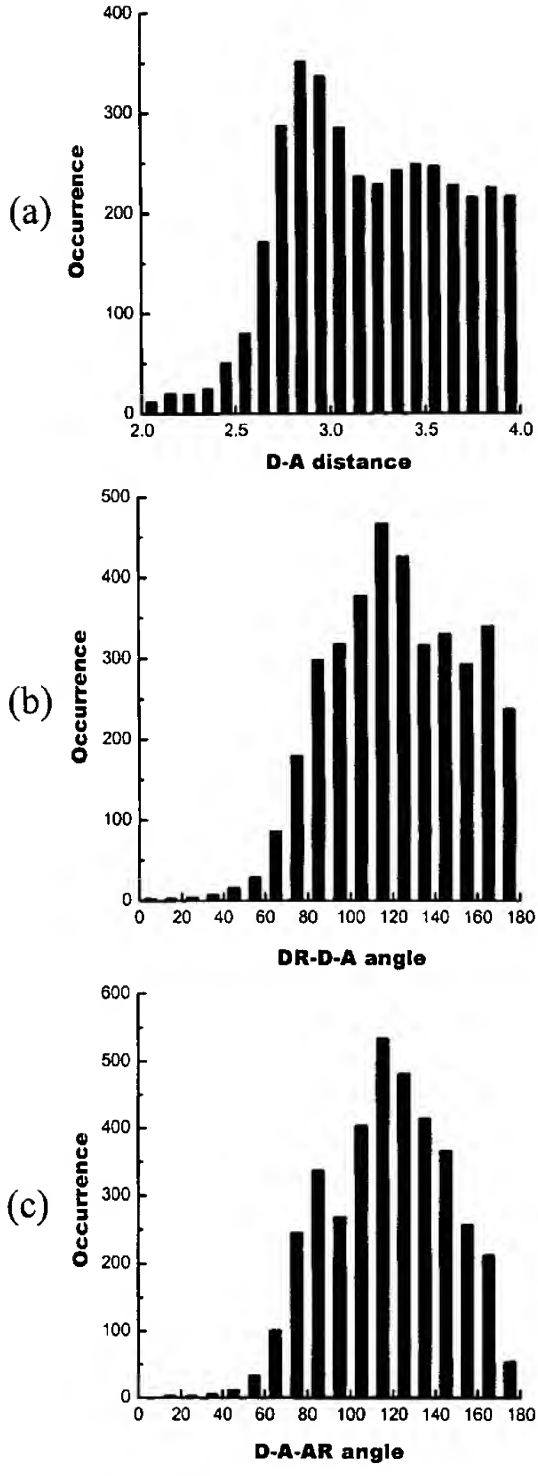


Figure 2. Distribution of the three geometric parameters in a hydrogen bond observed in the training set: (a) the donor-acceptor distance (in angstroms); (b) the DR-D-A angle (in degrees); (c) the D-A-AR angle (in degrees).

form should not exceed the number of lone pairs on that acceptor atom. If an atom could be a donor and an acceptor at the same time, such as the oxygen atom in a hydroxyl group, both rules apply.

As implied above, charged and neutral hydrogen bonds are not treated separately in our scoring functions since we find that the improvement of *our scoring functions* in the training set regression is totally negligible by separating them.

In some cases, metal ions are found inside the binding site of the protein. They may form coordinated bonds with atoms with lone pairs in the ligand and thus also contribute to the binding affinity. We include this kind of interaction in the hydrogen bonding term since it is the same as hydrogen bonding in nature, i.e. Lewis acid-base pair. Note that technically we define metal ions as 'donor' so that the metal-ligand coordinated bonds are calculated with exactly the same distance and angular functions of hydrogen bonding.

(4) Deformation effect. Upon binding, both the ligand and the protein will be constrained in conformation as compared to their free states in solvent. This will raise adverse entropic changes, which is a negative effect that must be overcome during the binding process. In other empirical scoring functions, the deformation effect of the ligand is often estimated by counting the number of rotatable bonds (rotors) that become frozen during the binding process, assuming that each rotor is associated with a discrete number of low-energy conformations and thus a certain amount of conformational entropy. If there are more than one rotor in the ligand, their contributions are assumed to be additive. This assumption is reasonable when all the rotors are isolated and free to rotate, so the low-energy conformations associated with each rotor will multiply to build up the entire conformational space. However, when two or more rotors cross, apparently this assumption is no longer valid because now the rotation of any of them will interfere with the others and this will result in a reduction in the total number of possible low-energy conformations (Figure 3). Therefore, simply counting the number of rotors often overestimates the conformational flexibility of certain kinds of molecules, such as oligo-peptides.

In our algorithm, rotor is defined as acyclic sp^3-sp^3 or sp^3-sp^2 single bond between two non-hydrogen atoms. Terminal groups, such as $-CH_3$, $-NH_2$, $-OH$, and $-X$ (X = F, Cl, Br, I), whose rotation does not produce any new conformation of heavy atoms are not counted as rotors. The potential flexibility of cyclic portions of the ligand is ignored. The deformation

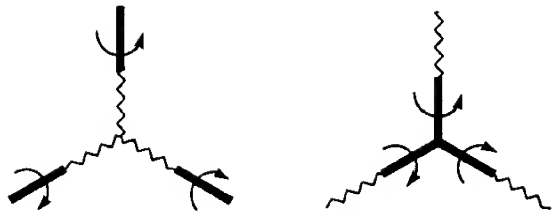


Figure 3. Illustration of 'isolated' (on the left) and 'crossed' rotors (on the right).

effect of the ligand is then expressed as the contribution of all the rotors with proper weight factors. For the convenience of computation, rotors are counted by summing the share of each ligand atom:

$$\mathrm{RT} = \sum_{i}^{ligand} \mathrm{RT}_i, \tag{5}$$

where $\mathrm{RT}_i = 0$ if atom i is not involved in any rotor; $\mathrm{RT}_i = 0.5$ if atom i is involved in one rotor; and $\mathrm{RT}_i = 1.0$ if atom i is involved in two rotors. However, if atom i is involved in more than two rotors, then $\mathrm{RT}_i = 0.5$. Note that, according to the conventional rotor-counting algorithm, RT_i should be 1.5 (in three rotors) or 2.0 (in four rotors) in this case. This reduction in the RT_i value reflects our consideration for offsetting the overestimation of conformational flexibility in the conventional algorithm. Although very crude, we found that our algorithm improves the accuracy of our scoring functions.

In our algorithm, the deformation effect of the protein is neglected. We have attempted to count the number of rotors presented in the side chains of the binding site residues and include it as a term in our scoring functions. However, such attempt did not improve the result. It is not surprising since the side chains of the binding site residues are generally immobilized even in the unbound state due to the stacking of neighboring residues. A more reasonable algorithm needs to be developed to account for the flexibility of the protein.

(5) Hydrophobic effect. Binding of the ligand and the protein is accompanied by the desolvation process that undergoes changes in entropy as well as in enthalpy. One of the results is that non-polar groups tend to favor each other, which is also referred to as 'hydrophobic effect'. This effect is very difficult for accurate characterization since it involves complicated ligand-water, protein-water, and water-water interactions before and after binding. Different algorithms have been used in other empirical scoring functions to calculate this term. We have implemented three representative algorithms in our scoring functions.

(i) Hydrophobic surface algorithm. The hydrophobic effect is assumed to be proportional to the buried hydrophobic surface of the ligand (Equation 6). This algorithm was adopted by Bohm's scoring function [27]. It should be pointed out that technically there are several types of molecular surfaces. Here we choose to use the solvent-accessible surface (SAS).

$$\mathrm{HS} = \sum_{i}^{ligand} \mathrm{SAS}_i. \tag{6}$$

The radius of the solvent probe is set to 1.5 Å. The solvent-accessible surface of the ligand is represented by evenly distributed dots in a spacing of 0.5 Å. Numerical integration is used to calculate the surface area. The surface areas of hydrogen atoms are attributed to their root atoms. Any part of the ligand surface is considered buried if it penetrates into the solvent-accessible surface of the protein. Note that only hydrophobic atoms are considered in Equation 6. The total amount of buried surface area is expressed in square Angstrom.

(ii) Hydrophobic contact algorithm. The hydrophobic effect is calculated by summing up the hydrophobic atom pairs formed between the ligand and the protein. This algorithm was adopted by ChemScore [30]. In our algorithm, it is calculated as:

$$\mathrm{HC} = \sum_{i}^{ligand} \sum_{j}^{protein} f\left(d_{ij}\right), \tag{7}$$

where

$$\begin{aligned} f(d) &= 1.0 && d \le d_0 + 0.5\ \text{Å} \\ &= (1/1.5) \times (d_0 + 2.0 - d) && d_0 + 0.5\ \text{Å} < d \le d_0 + 2.0\ \text{Å} \\ &= 0.0 && d > d_0 + 2.0\text{Å}. \end{aligned}$$

This distance function reflects the intuition that the strength of 'hydrophobic interaction' will reach the maximum when two hydrophobic atoms form van der Waals contact and diminish gradually with the increase in the inter-atomic distance. We find that this distance function needs to be fairly long-ranged in order to work well.

(iii) Hydrophobic matching algorithm. This algorithm was adopted by SCORE [32]. According to this method, different parts of the ligand sense the protein differently because of the heterogeneous nature of the binding site. If a hydrophobic ligand atom is placed at a hydrophobic site of the protein, then it is

expected to be favorable to the binding process. The overall hydrophobic matching between the ligand and the protein is calculated as:

$$\mathrm{HM} = \sum_{i}^{ligand} \log P_i \times \mathrm{HM}_i. \quad (8)$$

Here HM_i is an indicator variable. It is set to 1 if a hydrophobic atom i is placed in a hydrophobic environment; otherwise it is set to 0. $\mathrm{Log}P_i$ refers to the hydrophobic scale of atom i, which is the contribution of atom i to the n-octanol/water partition coefficient (Log P) of the molecule. In our algorithm, the hydrophobic scales for all kinds of atoms are cited from XLOGP2 [36]. They are introduced as weight factors here to ensure that more hydrophobic atoms contribute more to the hydrophobic effect. The 'environment' of a given ligand atom is defined to consist of all the atoms on the protein which are within 6 Å from the ligand atom. The hydrophobicity of the environment is determined by summing up the hydrophobic scales of all its member atoms. Our investigation of the training set shows that the average hydrophobicity of an environment surrounding a hydrophobic ligand atom is -0.50 log$P$ units. Therefore, in our algorithm an environment is defined as hydrophobic if its hydrophobicity is greater than -0.50 log$P$ units.

Finally, we summarize our scoring functions below. The binding affinity of a given protein-ligand complex, as expressed in pK_d unit, is calculated by summing up all the terms described above. Since three different algorithms for modeling the hydrophobic effect have been implemented, we have three resulting scoring functions:

$$\begin{aligned} pK_{d,1} = {} & C_{0,1} + C_{\mathrm{VDW},1} \times \mathrm{VDW} \\ & + C_{\mathrm{H-bond},1} \times \mathrm{HB} \\ & + C_{\mathrm{rotor},1} \times \mathrm{RT} \\ & + C_{\mathrm{hydrophobic},1} \times \mathrm{HS}, \end{aligned} \quad (9)$$

$$\begin{aligned} pK_{d,2} = {} & C_{0,2} + C_{\mathrm{VDW},2} \times \mathrm{VDW} \\ & + C_{\mathrm{H-bond},2} \times \mathrm{HB} \\ & + C_{\mathrm{rotor},2} \times \mathrm{RT} \\ & + C_{\mathrm{hydrophobic},2} \times \mathrm{HC}, \end{aligned} \quad (10)$$

$$\begin{aligned} pK_{d,3} = {} & C_{0,3} + C_{\mathrm{VDW},3} \times \mathrm{VDW} \\ & + C_{\mathrm{H-bond},3} \times \mathrm{HB} \\ & + C_{\mathrm{rotor},3} \times \mathrm{RT} \\ & + C_{\mathrm{hydrophobic},3} \times \mathrm{HM}. \end{aligned} \quad (11)$$

It should be emphasized that, except for the hydrophobic effect term, all the other terms in these three scoring functions are calculated using identical algorithms. The consensus scoring function, which is named as X-CSCORE, is the arithmetical average of Equations 9–11:

$$\mathrm{X-CSCORE} = \left(pK_{d,1} + pK_{d,2} + pK_{d,3}\right)/3 \quad (12)$$

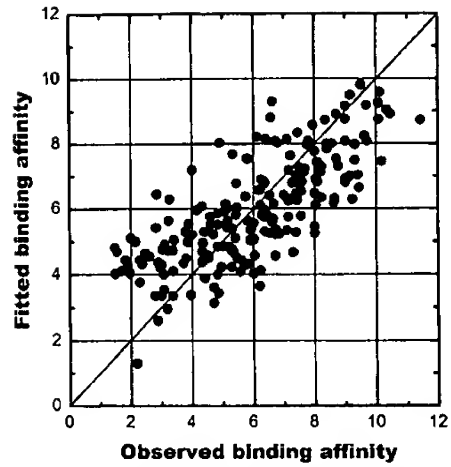


Figure 4. Correlation between the observed binding affinities of the 200 protein-ligand complexes in the training set and the fitted values given by X-CSCORE (in pK_d units).

Regression analyses

Coefficients before each term in Equations 9–11 are derived through standard least-square multivariate regression analyses of the training set. They are listed in Table 1 together with other related information. Correlation coefficients (r^2) and standard deviations (s) obtained from regression are listed in Table 2. The correlation between the observed binding affinities and the fitted values given by X-CSCORE is shown in Figure 4. Leave-one-out cross-validations are performed to judge the quality of the regression models. The resulting q^2 and s_{press} are listed in Table 2. Both the regression and the cross-validation are performed with the QSAR module in SYBYL.

Validation

(1) Test set. An independent test set is usually needed to validate a regression model. When constructing the training set, we deliberately separate all the complexes released by the Protein Data Bank after 1998 from the others. These complexes, 30 in total, are used as a test set in our study. A complete list of the test set can

Table 1. Regression models of Equations 9–11

Term	Coefficient[a]	Mean value[b]	Contribution fraction[c]
(Equation 9)			
VDW	-2.01×10^{-3} ($\pm$ 1.81×10^{-3})	-6.00×10^{2}	16.5%
H-Bond	0.307 ($\pm$ 0.137)	4.21	19.8 %
Rotor	-0.159 ($\pm$0.079)	7.28	25.3 %
Hydrophobic surface	7.10×10^{-3} ($\pm 2.50 \times 10^{-3}$)	2.74×10^{2} Å^2	38.4%
Constant	2.69 ($\pm$ 0.66)	-	-
(Equation 10)			
VDW	-0.96×10^{-3} ($\pm$ 1.91×10^{-3})	-6.00×10^{2}	8.6%
H-Bond	0.412 ($\pm$ 0.149)	4.21	29.4%
Rotor	-0.100 ($\pm$ 0.074)	7.28	17.5%
Hydrophobic contact	3.73×10^{-2} ($\pm$ 1.12×10^{-2})	43.1	44.5%
Constant	2.78 ($\pm$ 0.65)	-	-
(Equation 11)			
VDW	-2.14×10^{-3} ($\pm$ 1.65×10^{-3})	-6.00×10^{2}	16.4%
H-Bond	0.311 ($\pm$ 0.131)	4.21	18.8%
Rotor	-0.169 ($\pm$ 0.078)	7.28	25.2%
Hydrophobic matching	0.602 ($\pm$ 0.159)	2.51	39.6%
Constant	3.10 ($\pm$ 0.65)	-	-

[a] All coefficients are presented in pK_d units. They can be converted into binding free energies at 298 K in kcal/mol by multiplying a factor of -1.36. The values in brackets are 95% confidence intervals in regression.
[b] Mean values of each term are calculated over the entire training set.
[c] Contribution fractions are calculated by using the QSAR/PLS module in SYBYL.

Table 2. Statistical results of Equations 9–11 and X-CSCORE

	Equation 9	Equation 10	Equation 11	X-CSCORE
R^2	0.504	0.546	0.571	0.591
S^a	1.58	1.53	1.43	1.47
$F(4, 195)$	49.6	58.7	70.4	-
Q^2	0.480	0.522	0.551	-
S_{press}	1.62	1.57	1.47	-
R^2_{pred}	0.318	0.319	0.249	0.356
S_{pred}	1.51	1.61	1.63	1.58

[a] All the standard deviations, including S, S_{press} and S_{pred}, are presented in pK_d units. They can be converted into binding free energies at 298 K in kcal/mol by multiplying a factor of -1.36.

be found in the *supplementary material* section in this paper.

All the scoring functions, including Equations 9–12, are used to predict the binding affinities of the 30 protein-ligand complexes in the test set. The root-mean-squared deviation (s_{pred}) is used to measure the quality of prediction:

$$s_{\text{pred}} = \sqrt{\sum \left(pK_{d,\text{pred}} - pK_{d,\text{obs}}\right)^2 / (N-1)}. \quad (13)$$

The statistical results are shown in Table 2. The correlation between the experimentally observed binding affinities and the predicted values given by X-CSCORE is shown in Figure 5.

(2) Evolutionary regression. We have adopted an iterative regression procedure to further validate the internal consistency of our scoring functions, which was originally proposed in our previous work SCORE [32]. The central idea of this procedure, called evolutionary regression, is to test a given regression model with training sets of different sizes. In our study, this procedure starts from constructing a subset of 50 complexes which are randomly selected from the training set without duplication. This subset is used to perform multivariate regression and leave-one-out cross-validation for the scoring function under inves-

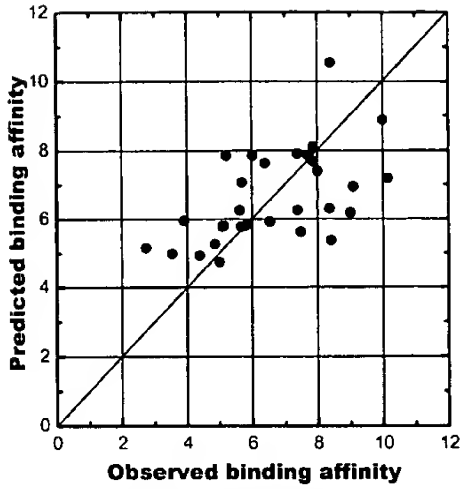


Figure 5. Correlation between the observed binding affinities of the 30 protein-ligand complexes in the test set and the predicted values given by X-CSCORE (in pK_d units).

tigation. All the regression results, including r^2, s, q^2, s_{press}, and the coefficients for each term in the scoring function, are recorded. This regression model is then used to predict the K_d values of the test set. The resulting r^2_{pred} and s_{pred} are also recorded. Since the subset is constructed randomly, the entire procedure, i.e. construction of the subset, multivariate regression, cross-validation, and calculation of the test set, is repeated for 10 times to reduce the noises in all the statistical results. Only the averaged results are used for analysis. At the next step, the size of the subset is increased by 10, and the regression model is re-evaluated with this new subset. This procedure is repeated until the size of the subset reached the full size of the training set. We have performed evolutionary regression for Equations 9–11. The standard deviations observed during the evolutionary regression procedure of Equations 9–11 are shown in Figure 6a–c, respectively.

(3) Molecular docking. We have also tested the performance of X-CSCORE in molecular docking experiments. We select 10 samples from the training set, including the L-arabinose binding protein/L-arabinose complex (PDB code 1ABE), the alcohol dehydrogenase/CNAD complex (PDB code 1ADB), the adenosine deaminase/DAA complex (PDB code 1ADD), the cytidine deaminase/uridine complex (PDB code 1AF2), the maltodextrin binding protein/maltose complex (PDB code 1ANF), the carboxypeptidase A/L-benzylsuccinate complex (PDB code 1CBX), the antibody DB3/progesterone analogue complex

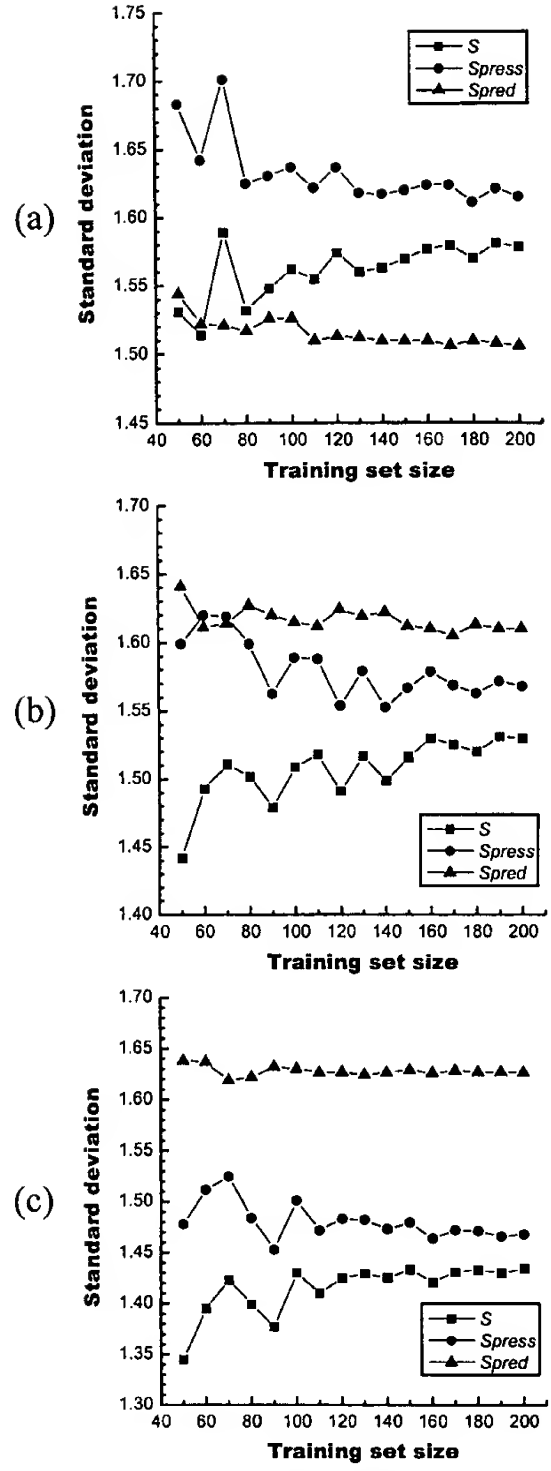
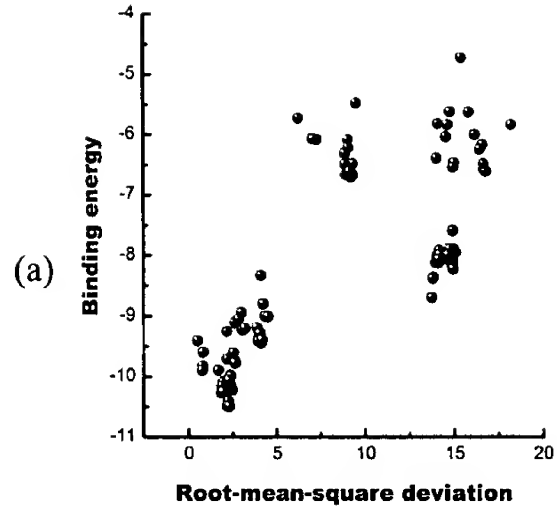


Figure 6. Standard deviations (in pK_d units) observed in the evolutionary regression procedure. (a) Equation 9; (b) Equation 10; (c) Equation 11.



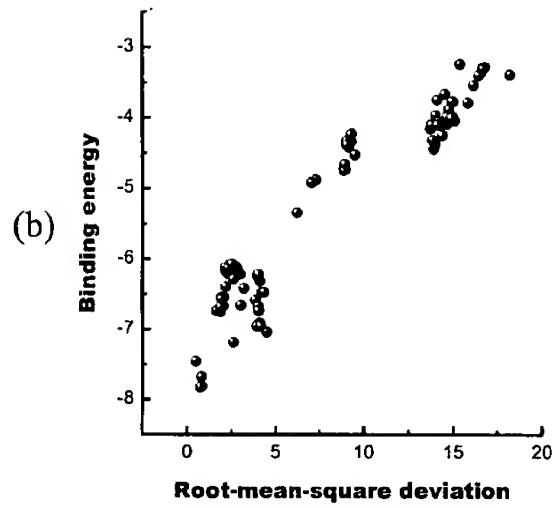
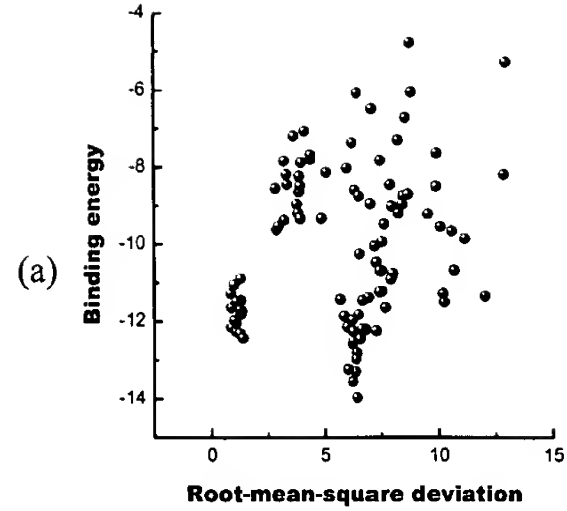


Figure 7. Relationship between the RMSD values (Å) and the binding energies (kcal/mol) of 100 conformations of L-benzylsuccinate in complex with carboxypeptidase A (PDB code 1CBX). (a) Binding energies calculated by AutoDock. (b) Binding energies calculated by X-CSCORE.



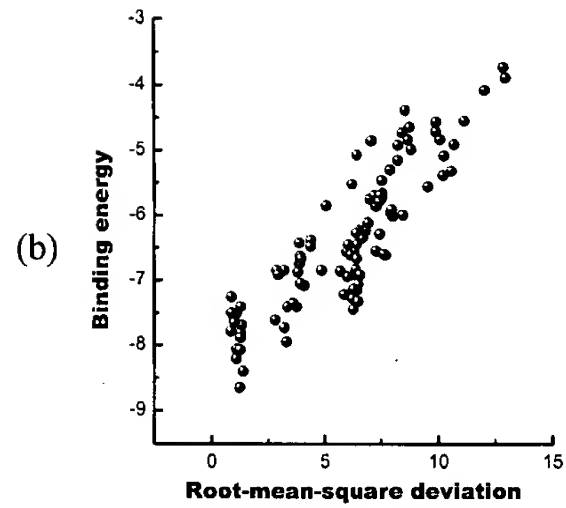


Figure 8. Relationship between the RMSD values (Å) and the binding energies (kcal/mol) of 100 conformations of folate in complex with dihydrofolate reductase (PDB code 1DHF). (a) Binding energies calculated by AutoDock. (b) Binding energies calculated by X-CSCORE.

(PDB code 1DBM), the dihydrofolate reductase/folate complex (PDB code 1DHF), the glutathione S-transferase/glutathione complex (PDB code 1GST), and the HIV-1 protease/VX-478 complex (PDB code 1HPV). The selection of these 10 samples emphasizes the diversity of the ligands and the proteins. For each complex, the AutoDock 3.0 program [8] is employed to perform a molecular docking run. In each case, the experimentally determined complex structure is always used as the starting point. The ligand is treated flexible while the protein is kept rigid. The searching steps in the conformational sampling for translation, quaternion, and torsion are set to 0.5 Å, 15° and 15°, respectively. Fifty thousand genetic algorithm generations are run with a population of 100 conformations. The final 100 best-scored conformations are saved and their root-mean-squared deviations (RMSD), as calculated by using the observed bound conformation as the reference, are recorded. Then the binding

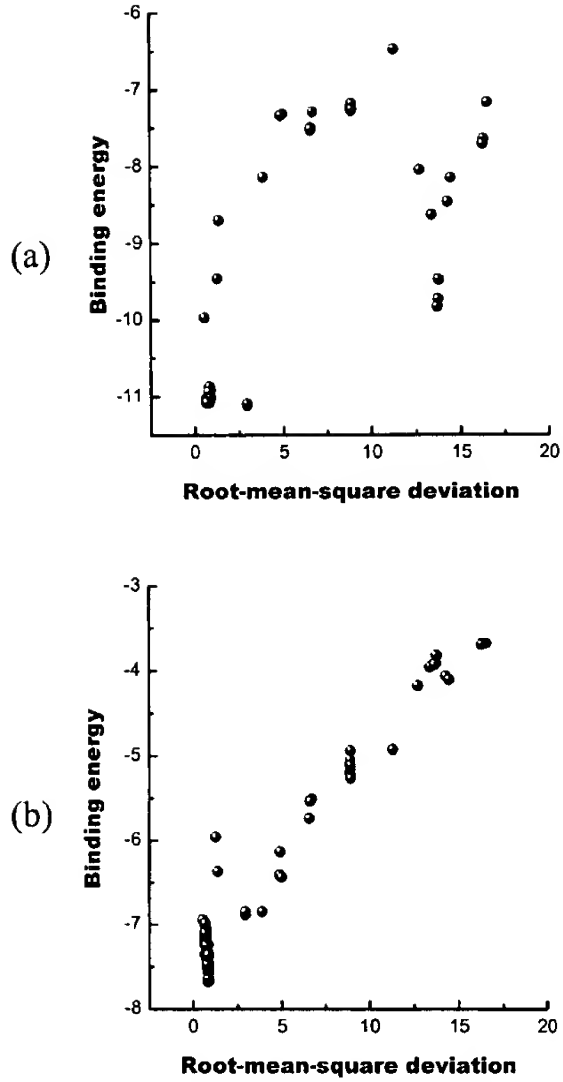


Figure 9. Relationship between the RMSD values (Å) and the binding energies (kcal/mol) of 100 conformations of 1-deaza-adenosine in complex with adenosine deaminase (PDB code 1ADD). (a) Binding energies calculated by AutoDock. (b) Binding energies calculated by X-CSCORE.

energies of these conformations are re-calculated by X-CSCORE. RMSD values of the best-scored conformations, picked by AutoDock and X-CSCORE, of all 10 complexes are summarized in Table 3. For complex 1CBX, 1DHF, and 1ADD, RMSD values of the final 100 conformations are plotted against the binding energies of these conformations in Figures 7–9, respectively.

Program description

We have developed a program, X-CSCORE, to implement the three scoring functions described by Equations 9–11. Here 'CSCORE' means consensus scoring; while the prefix 'X' indicates that it is part of our in-house drug design toolkit X-TOOL. This program is written in ANSI C++ and has been tested on UNIX and LINUX platforms. The required inputs include the three-dimensional structure of the protein in PDB format and the pre-docked ligand molecule(s) in MOL2 format. The user is allowed to enable or disable any of the three scoring functions in computation and the final predicted binding affinities are based on the arithmetic average of all the enabled scoring functions. If all three scoring functions are enabled, typically this program is able to process around 10 000 ligand molecules for a given protein target in an hour on a SGI O2/R5000/180MHz workstation.

Discussion

Accuracy and robustness

As shown in Table 2, Equations 9–11 are able to reproduce the binding affinities of the entire training set with standard deviations (s) of 1.58, 1.53 and 1.43 pK_d units, respectively. Their standard deviations in leave-one-out cross-validation (s_{press}) are at the same level, which are 1.62, 1.57 and 1.47 pK_d units, respectively. More importantly, these scoring functions perform almost equally well for the independent test set: the standard deviations in the predicted binding affinities (s_{pred}) are 1.51, 1.61 and 1.63 pK_d units, respectively. These values correspond to 2.1–2.2 kcal/mol in binding free energy at room temperature. Considering the diversity of the test set, such accuracy in binding affinity prediction is encouraging. Compared to other existing empirical scoring functions, our scoring functions have achieved better or comparable statistical results. If taking the Fisher significant ratio (F) as an objective criterion for comparing different regression models, the values are 32.1, 44.5 and 57.8 for Bohm's scoring function [27], ChemScore [30], and SCORE [32], respectively. In comparison, the F values of our scoring functions are 49.6, 58.7 and 70.4 for Equations 9–11, respectively.

When building a regression model, over-fitting of the regression equation should be avoided because it may fail to give reasonable predictions for samples

Table 3. Results from the molecular docking studies of 10 protein-ligand complexes

PDB code	Resolution (Å)	RMSD (Å)[a]		pK_d	
		AutoDock	X-CSCORE	Exp.[b]	X-CSCORE[c]
1ABE	1.7	0.62	0.73	6.52	5.14 (5.25)
1ADB	2.4	2.74	2.74	8.40	6.71 (8.01)
1ADD	2.4	2.93	0.83	6.74	5.63 (5.36)
1AF2	2.3	0.88	0.88	3.10	5.26 (4.90)
1ANF	1.67	0.60	0.54	5.46	6.16 (6.03)
1CBX	2.0	2.30	0.77	6.35	5.74 (5.74)
1DBM	2.7	1.31	1.13	9.44	6.84 (6.65)
1DHF	2.3	6.44	1.24	7.40	6.34 (5.27)
1GST	2.2	0.74	1.21	4.68	5.92 (5.21)
1HPV	1.9	1.73	1.16	9.22	6.47 (6.28)

[a]RMSD value of the best scored conformation in reference to the observed bound conformation.
[b]Experimentally determined binding affinity.
[c]Calculated binding affinity for the best scored conformation. The values in brackets are the calculated one for the observed bound conformation.

outside the training set. For this reason, only a minimal number of adjustable parameters are included in our scoring functions to achieve maximal N/M ratio in regression analysis. For example, we do not assign additional weighting factors to different types of atoms when calculating the van der Waals interaction. When calculating the hydrogen bonding, we do not differentiate charged and neutral hydrogen bonds. No differentiation in aliphatic and aromatic atoms was made in calculating the hydrophobic effect. Besides the regression constant, there are only four coefficients in each of our scoring functions. As shown in Table 1, they are all significant in regression analysis. Here the van der Waals interaction term in Equation 10 seems to be an exception, which contributes only a relative small fraction. However, it is not surprising since the hydrophobic effect term in Equation 10 is also calculated by counting atom pairs, therefore it overlaps with the van der Waals term partially and 'grabs' some contributions from the van der Waals term.

The N/M ratio issue deserves a little more discussion. It is reasonable to expect that statistically converged results can only be obtained by using a large training set. But how large is large? What is the proper size of the training set for deriving an empirical scoring function like ours? To answer this question, we have adopted the evolutionary regression procedure to look for the answer. The idea of evolutionary regression is to test a given regression model with training sets in different sizes and monitor the quality of the regression model during this procedure. Several trends can be seen in the evolutionary regression experiments of Equations 9–11 (Figure 6). (i) The standard deviation in the whole set fitting (s) gradually increases when the training set grows larger. This can be understood because the scoring function under regression is kept fixed during the whole procedure. A larger training set represents more complexity and thus is more difficult to reconcile. (ii) The predictive ability of the regression model, as indicated by the standard deviations in leave-one-out cross-validation (s_{press}) and test set computation (s_{pred}), is gradually improved when the training set grows larger. This indicates that a larger training set indeed helps our scoring functions achieve better predictive ability. (iii) When the training set is relatively small, the regression model is generally unstable. The final regression model depends very sensitively on the contents of the training set, which may lead to chance correlation in regression and poor predictive ability. When the training set grows larger, the regression model becomes more stable and tends to converge to a certain level. As suggested by our evolutionary regression experiments, a training set containing at least 160 samples is required to derive a stable empirical scoring function with four terms, i.e. a minimal N/M ratio of 40. Unfortunately, the N/M ratios of other existing empirical scoring functions are generally much lower than this, e.g. LUDI (N/M = 45/6 = 9) [27], ChemScore (N/M = 82/4 = 20) [30], and SCORE (N/M = 170/10 = 17) [32]. In our case, the N/M ratio is 200/4 = 50. Therefore, we believe our scoring functions are, if not much more accurate, more

robust in binding affinity prediction for a wider range of protein-ligand complexes.

Consensus scoring

A unique feature of our study is that three different algorithms have been implemented for modeling the hydrophobic effect. As described in the Methods section, hydrophobic effect is calculated either by the buried solvent-accessible molecular surface (Equation 9), or by the number of hydrophobic contacts between the protein and the ligand (Equation 10), or by the hydrophobic matching of the ligand with the binding site (Equation 11). All three algorithms are conceptually acceptable and actually they represent three typical algorithms adopted by empirical scoring functions for modeling the hydrophobic effect. However, it is not a good idea to include all three terms together in *one* scoring function since they account for the same effect and thus are highly correlated to each other. Therefore, they have to be accommodated in three scoring functions. As indicated by our regression results (Table 2), all three scoring functions perform reasonably well and are basically comparable to each other. However, since these three algorithms utilize different geometric features of the given protein-ligand complex structure in computation, their results differ. We have found that, for 40.0% of the samples in the training set, the difference between the lowest and the highest calculated binding affinity by these three scoring functions is less than 0.50 pK_d units; for 40.5% of the samples, the difference is between 0.50 and 1.00 pK_d units; while for the remaining 19.5% of the samples, the difference is larger than 1.00 pK_d units. One can see that such difference is not trivial at all in many cases. Conceivably, if one can predict which scoring function will be the best for a given protein-ligand complex, the accuracy in binding affinity prediction will be improved greatly. Indeed, if the experimental values are correlated to the best fitted values (each of them is chosen from three hits), the standard deviation in the training set fitting will drop by half to about 0.7 pK_d units. We have attempted to find out which scoring function may perform better for certain classes of ligands or families of proteins. Unfortunately this attempt ended without much success.

Based on the fact that there is no reason to bias towards any one of the three scoring functions, we simply combine them together (Equation 12). This practice is consistent with the idea of consensus scoring which has been demonstrated to be an effective way of improving the hit-rates in virtual database screening [33]. As shown in Table 2, the performance of a single scoring function may vary and is not predictable. For example, among the three scoring functions, Equation 9 is the worst one for the training set but the best one for the test set. In contrast, Equation 11 is the best one for the training set but the worst one for the test set. By averaging these scoring functions, i.e. X-CSCORE, the result is not always the closest one to the true value (in fact it is always between the best one and the worst one). However, the advantages are: (i) it provides a clear indication of what level of accuracy these three scoring functions can achieve. Obtaining a converged result in binding affinity prediction is certainly important for structure-based drug design practice; and (ii) large errors in binding affinity prediction can be reduced. Recently we have pointed out that the nature of consensus scoring is multiple sampling [37]. By applying multiple scoring functions in combination, the positive and the negative errors have a chance to cancel each other and that is why consensus scoring generally performs better than any single scoring procedure.

Application to molecular docking

Our scoring function is developed primarily for estimating the binding affinity of a given complex with known structure ('scoring'). We also expect it to be useful for identifying the correct 'pose' of a ligand to its receptor ('docking'). Although some disputes still exist in whether 'docking' or 'scoring' should use the same type of function, we believe that ideally a 'scoring' function should also be able to serve as a 'docking' function. This is very important because in practice 'docking' and 'scoring' are often inseparable, such as in a virtual database screening study.

As described in the *Methods* section, we have investigated the potential application of X-CSCORE in molecular docking with 10 samples. Since we have not implemented this consensus scoring function into any molecular docking program directly, we employ the AutoDock program as a tool to generate possible bound conformations of the given ligand. All the conformations are then re-evaluated by X-CSCORE. RMSD values of the best scored conformations of these 10 protein-ligand complexes are listed in Table 3, where the results of X-CSCORE and the force field calculation in AutoDock are compared side by side. As one can see, if using the force field calculation in AutoDock as the scoring engine, 4 out

of the total 10 samples have RMSD values larger than 2.0 Å; while if using X-CSCORE as the scoring engine, only one sample, i.e. the alcohol dehydrogenase/CNAD complex (PDB code 1ADB), shows a RMSD value larger than 2.0 Å. In this case, we have checked all the 100 conformations generated by AutoDock and we found that the lowest RMSD value is 2.74 Å. This indicates that, with the parameters we were using, AutoDock has not generated any conformation close enough to the observed one. In fact, X-CSCORE predicts a much higher pK_d value of 8.01 for the observed one. The RMSD versus energy relationships observed in our docking tests for the Carboxypeptidase A/L-benzylsuccinate complex (PDB code 1CBX), the Dihydrofolate reductase/folate complex (PDB code 1DHF), and the adenosine deaminase/DAA complex (PDB code 1ADD) are shown in Figures 7–9, respectively. For these three samples, the best RMSD values given by AutoDock are 2.30 Å, 6.44 Å and 2.93 Å; while the corresponding ones given by X-CSCORE are 0.77 Å, 1.24 Å and 0.83 Å. It is very interesting to notice that, in the case of 1DHF, AutoDock has apparently chosen a wrong class of conformations while the correct one is somehow scored about 2 kcal/mol higher. In contrast, X-CSCORE has no problem in identifying the correct conformation.

It is very encouraging that our scoring functions are also applicable to molecular docking. Our scoring functions have all the necessary elements that correspond to the non-covalent interactions in a conventional force field, such as the van der Waals interaction and the electrostatic interaction (replaced by the hydrogen bonding term in our scoring functions). Besides that, our scoring functions also consider the hydrophobic effect and thus provide a better estimation of binding free energies. This is suggested in Figures 7b, 8b and 9b. In these cases, there is always a clear correlation between the RMSD values of the conformations and their binding energies calculated by X-CSCORE. Generally, the smaller is the RMSD value, the lower is the binding energy. The importance of this feature should not be underestimated. Molecular docking is a conformational sampling procedure which is performed on the potential energy surface defined by a certain scoring function. It is important that this potential energy surface does not contain a large number of false minima since such frustration will probably lead to poor convergence or wrong binding modes. The potential energy surface defined by an ideal scoring function should shape like a funnel, on which all the paths finally go down to the right position. As indicated by the RMSD versus energy relationships shown in Figures 7b, 8b and 9b, our consensus scoring function may have such an appealing feature. We expect that if a molecular docking program adopts our consensus scoring function as its scoring engine, its accuracy and efficiency in finding the correct bound structure will be improved considerably.

Considering that in practice our consensus scoring function will be applied in conjunction with molecular docking programs, it is highly desirable that all our scoring functions are able to tolerate at least a small amount of uncertainty in the input structure. For this reason, we have designed our scoring functions in such a way that they are not too sensitive to atomic coordinates. For example, we avoid the explicit use of hydrogen atoms in our algorithms. The reason is that predicting the position of a hydrogen atom precisely could be problematic when the hydrogen atom is bonded to a terminal rotatable group, such as a hydroxyl group. This uncertainty will lead to large deviation if hydrogen atoms have to be included explicitly in the calculation. Secondly, all the terms in our scoring functions are calculated with relatively large tolerances. For example, a 'softer' 8–4 equation is adopted in the van der Waals interaction term; loose criteria for distance and angular dependence are adopted in the hydrogen bonding term; long-distance cutoff is adopted in the hydrophobic effect terms. All these efforts are dedicated to emphasize on the overall fitness of the ligand to the binding site rather than trivial structural details. As shown in Table 3, by applying X-CSCORE, if a conformation is close to the reference conformation, then indeed it will get a score close to the one of the reference conformation.

Strength and weakness

Our scoring functions are developed to provide fast binding affinity estimations for a wide range of proteins and ligands. As demonstrated by the training set and the test set, the average accuracy of our consensus scoring function in calculating absolute binding free energies is approximately 2 kcal/mol. This level of accuracy is acceptable for structure-based lead discovery in which very accurate prediction of binding free energies may not be necessary, such as virtual database screening or *de novo* structure generation. The speed of our consensus scoring function is also perfectly suitable for such approaches.

We have implemented our scoring functions in a user-friendly program and have already applied it to several on-going structure-based drug design projects in our group. In these projects, large chemical databases are screened first by a standard docking program, such as DOCK, to pick out the top 10% compounds. These compounds are then re-evaluated by X-CSCORE. The best compounds selected by X-CSCORE, usually less than 0.1% of the original database, are then tested in biological assays. Very promising compounds have been identified since the application of this approach.

However, the accuracy of our consensus scoring function in binding affinity prediction is still not totally satisfactory: an error of 2 kcal/mol in binding free energy equals to approximately 50 folds in dissociation constant. Several drawbacks in our approach may have contributed to this inaccuracy. Firstly, since our scoring functions are derived from regression, they tend to characterize only the "common" interactions that are exhibited by a large population in the training set. Some other types of interactions, such as cation-π interaction and π-π stacking, are not included in our scoring functions simply because they have rare occurrences and thus do not contribute much to the regression model. It is thus expected that a general-type scoring function like ours could fail to give reasonable predictions when these types of interactions are playing an important role in protein-ligand binding. Secondly, there are also some factors which are common but we do not really have reasonable methods to take them into account. One example is the water molecules existing on the protein-ligand interface. Such water molecules are quite common and in some cases are thought to play an important role in the ligand binding. However, it remains unclear how to consider water molecules explicitly with an empirical scoring function. If water molecules need to be considered explicitly, maybe the entire algorithm for modeling the so-called 'hydrophobic effect' needs be replaced as well.

Our scoring functions also tend to give large positive errors for complexes with very low affinities and large negative errors for complexes with very high affinities (Figure 4). This phenomenon contributes to the significant positive intercept (~ 3 pK_d units) observed in all three scoring functions. Given the fact that most of the samples in the training set (80%) have pK_d values between 3.00 and 9.00, our scoring functions are calibrated better for binding affinities at this range. In fact, if only the samples within this affinity range are chosen to derive our scoring functions, the standard deviations in regression will drop to 1.2–1.3 pK_d units (~ 1.7 kcal/mol in binding free energy).

Another major problem is the quality of the training set. Ideally, each protein-ligand complex in the training set should have a known high-resolution three-dimensional structure together with a reliably measured binding affinity value accessible to the public. Obtaining protein-ligand complex structures is not a problem since the Protein Data Bank provides an excellent resource for such information. However, collecting the binding affinities for these complexes is a tedious job since they all scatter in various literatures. So far, no appreciable database for such information has been established. The training set used in our study is a compilation of the training sets published in others' work plus our own collections from the literature. Containing 200 samples, it is already the largest set published to date in an empirical scoring function approach. As demonstrated in our evolutionary regression test, the size of this training set is sufficient for calibrating our scoring functions. However, the binding affinity data presented in this training set still need careful examination because a large portion of them are cited directly from others' work without further confirmation. Besides, some of the dissociation constants could have been measured under different experimental conditions, such as PH level, temperature, and salt concentration. The uncertainties in the binding affinity data have certainly placed an intrinsic limit on the accuracy of our scoring functions.

It should be mentioned that all the drawbacks we have discussed above are shared by other empirical scoring functions as well. Despite of all these drawbacks, empirical scoring functions remain a valuable and indispensable means for structure-based drug design. Constructing a better training set will not be a problem in the future because more and more structural and binding affinity data are becoming available. We are also optimistic that better algorithms will appear to account for the binding process. All these efforts will lead to a substantial improvement in the performance of future empirical scoring functions.

Conclusion

We have developed a consensus empirical scoring function, X-CSCORE, for estimating the binding affinity of a given protein-ligand complex with a known three-dimensional structure. The framework

of our study is very similar to Böhm's pioneering work. However, we have presented our works on designing better algorithms for the contributing terms and calibrating the scoring functions against a larger training set. As shown in this paper, our consensus scoring function is able to predict the binding free energies with an average accuracy of approximately 2 kcal/mol. Its potential application to molecular docking is demonstrated with a number of protein-ligand complexes. When compared to the conventional force field calculation, X-CSCORE performs considerably better in identifying the correct bound conformations. Considering the reasonable accuracy, the wide applicability, and the respectable speed, we expect that X-CSCORE will become a valuable tool for structure-based drug design.

Supplementary material

Tables of the training set (200 protein-ligand complexes) and the test set (30 protein-ligand complexes). The program, X-CSCORE, is available by contacting the authors.

Acknowledgements

This work is financially supported by the Cap CURE Foundation (2001 Young Investigator Award to Dr Renxiao Wang) and the Department of Defense (Grant No. DOD DAMP17-93-V-3018 to Dr Shaomeng Wang). The authors are grateful to Dr John B. O. Mitchell at University of Cambridge for providing some of the binding affinity data used in this study. The authors are also grateful to Dr Chao-Yie Yang at University of Michigan Medical School for his many thoughtful suggestions.

References

1. Kuntz, I.D., *Science*, 257 (1992) 1078.
2. Greer, J., Erickson, J.W., Baldwin, J.J. and Varney, M.D., J. Med. Chem., 37 (1994) 1035.
3. Verlinde C.L.M.J. and Hol W.G.J., Structure, 2 (1994) 577.
4. Babine, R.E. and Bender, S.L., Chem. Rev., 97 (1997) 1359.
5. Gane, P.J. and Dean, P.M., Curr. Opin. Struct. Biol., 10 (2000) 401.
6. Walters, W.P., Stahl, M.T. and Murcko, M.A., Drug Discovery Today, 3 (1998) 160.
7. Makino, S. and Kuntz, I.D., J. Comp. Chem., 18 (1997) 1812.
8. Morris, G.M., Goodsell, D.S., Halliday, R., Huey, R., Hart, W.E., Belew, R.K. and Olson, A.J., J. Comput. Chem., 19 (1998) 1639.
9. Jones, G., Wilett, P., Glen, R.C., Leach, A.R. and Taylor, R., J. Mol. Biol., 267 (1997) 727.
10. Rarey, M., Kramer, B., Lengauer, T. and Klebe, G., J. Mol. Biol., 261 (1996) 470.
11. Böhm, H.J., Curr. Opin. Biotech., 7 (1996) 433.
12. Miranker, A. and Karplus, M., Proteins, 11 (1991) 29.
13. Böhm, H.J., J. Comput. Aid. Mol. Des., 6 (1992) 61.
14. Gillet, V., Johnson, P. and Mata, P., J. Comput. Aid. Mol. Des., 7 (1993) 127.
15. Clark, D.E., Frenkel, D. and Levy, S.A., J. Comput. Aid. Mol. Des., 5 (1995) 13.
16. Pearlman, D.A. and Murcko, M.A., J. Med. Chem., 39 (1996) 1651.
17. Wang, R., Gao, Y., Lai, L., *J. Mol. Model.*, 6(2000) 498-516.
18. Schneider, G., Lee, M.L., Stahl, M. and Schneider, P., J. Comput. Aid. Mol. Des., 14 (2000) 487.
19. Kollman, P.A., Curr. Opin. Struct. Biol., 4 (1994) 240.
20. Ajay and Murcko, M.A., J. Med. Chem., 38 (1995) 4953.
21. Tame, J.R.H., J. Comput. Aid. Mol. Des., 13 (1999) 99.
22. Goodford, P.J.A., J. Med. Chem., 28 (1985) 849.
23. Massova, I. and Kollman, P., Perspect. Drug Disc. Des., 18 (2000) 113.
24. Kollman, P., Chem. Rev., 7 (1993) 2395.
25. Aqvist, J., Medina, C. and Samuelsson, J.E., Protein Eng., 7 (1994) 385.
26. Carlson, H.A. and Jorgensen, W.L., J. Phys. Chem., 99 (1995) 10667.
27. Böhm, H.J., J. Comput. Aid. Mol. Des., 8 (1994) 243.
28. Jain, A.N., J. Comput. Aid. Mol. Des., 10 (1996) 427.
29. Head, R.D., Smythe, M.L., Oprea, T.I., Waller, C.L., Green, S.M. and Marshall, G.R., J. Am. Chem. Soc., 118 (1996) 3959.
30. Eldridge, M.D., Murray, C.W., Auton, T.R., Paolini, G.V. and Mee, R.P., J. Comput. Aid. Mol. Des., 11 (1997) 425.
31. Böhm, H.J., J. Comput. Aid. Mol. Des., 12 (1998) 309.
32. Wang, R., Gao, Y. and Lai, L., J. Mol. Model., 4 (1998) 379.
33. Charifson, P.S., Corkery, J.J., Murcko, M.A. and Walters, W.P., J. Med. Chem. 42 (1999) 5100.
34. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., Nucleic Acids Res., 28 (2000) 235, http://www.rcsb.org/pdb/.
35. SYBYL v6.2, Tripos Inc. St. Louis, MO, U.S.A. http://www.tripos.com/
36. Wang, R., Gao, Y. and Lai, L., Perspect. Drug Disc. Des., 19 (2000) 47.
37. Wang, R. and Wang, S., J. Chem. Inf. Comput. Sci., 41 (2001) 1422.